Tijdschrift voor Medisch Onderwijs (maart 2007) 26:59–63
DOI 10.1007/BF03056776

ONDERZOEK

De norm bij een didactische huisartsopleiderstoets: 'deed ik het goed genoeg?'

J.P. Lettinga · P.M. Boendermaker

Samenvatting Inleiding: Om de onderwijsvaardigheden van (huis)artsopleiders te toetsen, wordt in Groningen gebruik gemaakt van een didactische vaardigheidstoets met multipele stations: de PACT. Deelnemers wilden weten of ze (onderdelen van) de toets goed genoeg hadden gedaan. Onderzocht werd welke norm daarvoor geëigend is, passend bij het educatieve doel van de toets, en wat de consequentie zou zijn van het kiezen van een bepaalde norm.

Methode: Met behulp van een consensusmethode werden in elk station kernitems geïdentificeerd door 7 stafleden, 9 observatoren, 11 deelnemers en 10 simulatie-aios. Onderzocht werd wat het effect op 'alle-kernitems-voldoende/goed' percentages per station zou zijn bij de in de kernitems-consensusprocedure gehanteerde overeenstemming van minimaal 75%. Resultaten: Als gerekend wordt met een overeenstemmingsmaat van 75% blijkt dat van de 25 items in de vier stations in de toets er 14 als kernitem worden beschouwd. Toegepast op een recente toets betekent dit dat per station tussen de 56-78% van de deelnemers alle kernitems voldoende/goed deed.

Conclusie en beschouwing: Deze resultaten lijken de conclusie te rechtvaardigen dat het mogelijk is om voor de PACT per station een aanvaardbare norm te stellen, die recht doet aan het educatieve karakter van de toets. (Lettinga JP, Boendermaker PM. De norm bij een didactische huisartsopleiderstoets: 'deed ik het goed genoeg?'. Tijdschrift voor Medisch Onderwijs 2007;26(2):75-81.)

J.P. Lettinga (✉)
Dhr. J.P. Lettinga, student geneeskunde, Rijksuniversiteit Groningen.

Inleiding

De huisartsopleiding in Nederland duurt drie jaar. Het eerste en derde jaar van de opleiding van elke huisarts-in-opleiding (aios) bestaat uit een stage van twee jaar bij twee verschillende huisartsopleiders (bij elke opleider een jaar). Deze huisartsopleiders hebben dus twee functies: huisarts en opleider.

Conform de huidige onderwijskundige inzichten, die zelfgestuurd onderwijs stimuleren, is scholing op maat voor de huisartsopleiders in opkomst. Hiermee wordt beter tegemoet gekomen aan de individuele scholingsbehoefte van opleiders.

Om te weten welke lacunes – scholingsbehoefte – een huisartsopleider nog ervaart in zijn opleiderschap, is een vorm van toetsing behulpzaam. Zo is in de afgelopen jaren een, helaas nog beperkt, aantal toetsinstrumenten voor de huisartsopleider ontwikkeld. Diverse methoden van toetsing zijn daarbij denkbaar, zoals een visitatie, een vorm van zelftoetsing, het oordeel van de aios, een kennistoets of een didactische vaardighedentoets.[1] De didactische vaardighedentoets is een educatief bedoelde toetscarrousel met multipele stations, waarin gestandaardiseerde simulatiehuisartsen- in-opleiding meespelen, zoals simulatiepatiënten bij een medisch inhoudelijke vaardighedentoets. In deze toets worden de deelnemers door daartoe getrainde collega's geobserveerd met behulp van gestructureerde scorelijsten. De term 'objective structured teaching examination' (OSTE) werd voor deze vorm van toetsing van artsopleiders gekozen, vergelijkbaar met de OSCE bij de klinische vaardigheden, waarbij de 'c' staat voor 'clinical'.

Bij het Interuniversitair Centrum voor Huisartsopleiding (ICHO) in Vlaanderen is een dergelijke toets voor huisartsopleiders ontwikkeld: de multiple-station

teaching assessment test (MSTAT).[2-3] Bij de afdeling Huisartsopleiding in Groningen werd hierop voortgebouwd, hetgeen leidde tot de PACT: de 'physicians' assessment of competence in teaching'. In figuur 1 is de inhoud van deze toets per station weergeven, met de toetscriteria als items.

Bij deze toets ontbrak een norm. Bij het samenstellen van de toets was hierin nog niet voorzien. Dit werd als een gemis ervaren door de deelnemers (huisartsopleiders). Zij hadden behoefte aan een ijkpunt, waaruit zou blijken of ze een bepaalde didactische vaardigheid op een voldoende niveau beheersen. Dit leidde tot de volgende vraagstellingen:

1. Welke normstelling is voor de PACT geëigend voor het educatieve doel van de toets.
2. Wat is de consequentie voor de zak/slaag-verdeling van de toets bij het kiezen van een bepaalde norm?

Overzicht over toepassingen van de OSTE

Een literatuursearch resulteerde in vijf relevante verwijzingen op het gebied van het toetsen van artsopleiders met behulp van een OSTE (voor zoekstrategie: zie kader). Prislin et al. beschrijven in 1998 een onderzoek naar de toepasbaarheid van het gebruik van een OSTE om de onderwijsvaardigheden van huisartsdocenten te beoordelen.[4] Bij de scoring van de stations door staflidobservatoren bereikten drie van de acht stations een acceptabele intraclass correlatie. De waardering van de realiteit van de stations door staflidobservatoren was hoog. In de ogen van de deelnemers heeft een OSTE een bescheiden bruikbaarheid. De observatoren daarentegen vonden dat deze vorm van toetsing bruikbaar is voor de ontwikkeling van opleidersonderwijs voor artsdocenten.

In Medline zijn de volgende MeSeH termen gebruikt:

1. Family-practice-education

STATION 1
Leerplan maken

1-1 Leerbehoeften verhelderen
1-2 Concrete leerdoelen omschrijven
1-3 concreet leerplan opstellen
1-4 Afspraken maken voor uitvoering en opvolging

STATION 2
Observeren en feedback geven

2-1 Met de haio afspreken waarop de observatie zich gaat richten
2-2 De haio het woord geven
2-3 Eigen feedback geven
2-4 Gegeven feedback afchecken
2-5 Alternatieven bespreken
2-6 Maken van afspraken voor het nagaan van het effect van de feedback
2-7 Blijk geven van adequate analyse van de observatie

STATION 3
Tussentijdse evaluatie van het leerplan

3-1 Over uitvoering leerplan laten rapporteren
3-2 Voortgang waarderen
3-3 Exploreren van wat niet gelukt is en waarom
3-4 Weergeven van eigen indruk hierover
3-5 Eigen rol als opleider
3-6 Oplossingsgericht denken
3-7 Afspraken agenderen
3-8 Communicatie aspecten

STATION 4
Tussentijdse evaluatie van het leerplan

4-1 Beginniveau bepalen
4-2 Rekening houden met de leerstijl van de haio
4-3 Het onderwijzen van de vaardigheid in engere zin
4-4 Omgang met de patiënt
4-5 Feedback
4-6 Maken van afspraken

Figuur 1 Items per station.

2. Preceptor or ship
3. Teaching methods
4. Educational measurement

Andere databases, o.a. EMBASE en ERIC, hebben geen andere publicaties opgeleverd.

Bij het ICHO in Vlaanderen ontwikkelde Schol in 2000 een nieuwe vaardigheidstoets in stationsvorm om de didactische vaardigheden van huisartsopleiders vast te stellen.[2] De multiple-station teaching assessment test (MSTAT) blijkt een betrouwbaar, valide en acceptabel instrument. Vijf van de zeven stations hebben een goede interbeoordelaarsbetrouwbaarheid en de test is met name geschikt voor de screening van huisartsopleiders met betrekking tot hun onderwijsvaardigheden.[3]

Morrison et al. publiceerden in 2002 een artikel over de ontwikkeling van een OSTE voor residents as teachers (arts-assistenten in opleiding, die zelf docent zijn voor basisartsen in opleiding).[5] De interbeoordelaarsbetrouwbaarheid en de interne consistentie zijn hoog bij deze OSTE. Ook de inhoudsvaliditeit is hoog.

Uit onderzoek van Boendermaker (2003) blijkt dat ook de PACT een bruikbare toets is en geschikt als educatief toetsinstrument voor een aantal huisartsopleiderskenmerken. [6] De interne consistentie bleek acceptabel bij alle stations, de interbeoordelaarsbetrouwbaarheid was acceptabel in twee van de vier stations. De inhoudsvaliditeit is goed en de criteriumvaliditeit hoog.

In 2004 rapporteren Zabar et al. over de ontwikkeling, uitvoering en evaluatie van stations, waarin arts-assistenten worden getoetst op hun onderwijsvaardigheden.[7] Deze stations maken deel uit van een jaarlijkse OSCE voor arts-assistenten. De scoring op deze stations blijkt betrouwbaar en valide.

In geen van bovenstaande publicaties wordt over een mogelijke normstelling gesproken.

De eerste vraag die beantwoord moest worden was óf het zinvol is een norm te stellen voor een educatieve toets. Naar ons idee kan een educatief bedoelde toets als de PACT, die qua opzet en inhoud steeds gelijk is en die bedoeld is om vast te stellen wat al *goed* gaat en wat nog te verbeteren is, juist winnen aan educatieve waarde mét een norm, zodat vastgesteld wordt welke didactische vaardigheid al 'goed genoeg' gaat, oftewel voldoende wordt beheerst, en welke nog niet. Op dat laatste zal dan de scholing op maat zich moeten richten. Bovendien is het dan ook mogelijk om bij herdeelname te zien of de norm wél wordt gehaald.

Methode

Om tot een normstelling te komen, werd besloten een consensusprocedure te hanteren. [8] Aan alle deelnemers, simulatieaio's en observatoren die hebben meegewerkt tijdens een recente toetsafname, werden de scorelijsten van vier stations toegestuurd met de vraag om per item aan te geven of het in ieder geval voldoende of goed gescoord moet worden in dit station om de didactische vaardigheid op het hele station als voldoende te beschouwen, m.a.w. of het om een kernitem gaat. Tevens hebben we een aantal stafleden die betrokken zijn bij het opleidersonderwijs gevraagd om dit te doen. In deze consensusprocedure werden de resultaten van de eerste ronde per groep (deelnemers, observatoren, et cetera) aan de respondenten bekend gemaakt en werd gevraagd om met die informatie nog een keer aan te geven wat de kernitems zijn.

In veel consensusprocedures wordt een afkappunt van 75% consensus gehanteerd.[8] We hebben dit, in wezen arbitraire, uitgangspunt overgenomen. Daarbij werd vooraf beoogd dat, gezien het educatieve karakter van de toets, een uitkomst wenselijk zou zijn die per station groeimogelijkheden voor de deelnemers zichtbaar maakte, maar niet (te) demotiverend zou moeten zijn (b.v. dat *iedereen* met deze maatlat uiteindelijk *op alle stations* onvoldoende zou scoren).

Om te toetsen of met 75% overeenstemming een aanvaardbare verzameling kernitems ontstond, werd besloten met de resultaten van deze consensusprocedure de toetsscores van de meest recente opleiderstoets te benaderen. 'Aanvaardbaar' betekent in dit verband een zodanige slaag/zakverdeling dat meer mensen slagen dan zakken.

Resultaten

De respons na de tweede ronde varieert per groep van 59-88% (zie tabel 1). De voornaamste redenen voor de nonrespons waren ziekte en vakantie.

Tabel 1 Aantal aangeschrevenen en respons in de eerste en tweede ronde.

	Stafleden	Observatoren	Deelnemers	Simulatieaios
Aangeschreven	8	14	18	17
Respons I	7	10	14	14
Respons II	7	9	11	10

Tabel 2 Percentage dat het item als kernitem beoordeelt in station 1.

	Stafleden	Observatoren	Deelnemers	Simulatieaios	Gem. 4 groepen
Item 1	100	100	100	90	98
Item 2	86	100	91	67	86
Item 3	86	89	91	70	84
Item 4	71	56	73	30	57

Tabel 3 Percentage dat het item als kernitem beoordeelt in station 2.

	Stafleden	Observatoren	Deelnemers	Simulatieaios	Gem. 4 groepen
Item 1	71	100	100	33	76
Item 2	86	78	100	100	91
Item 3	100	100	100	100	100
Item 4	86	89	55	60	72
Item 5	43	11	55	90	50
Item 6	14	33	64	10	30
Item 7	57	0	36	20	28

Tabel 4 Percentage dat het item als kernitem beoordeelt in station 3.

	Stafleden	Observatoren	Deelnemers	Simulatieaios	Gem. 4 groepen
Item 1	100	100	91	100	98
Item 2	100	89	64	50	76
Item 3	100	100	100	100	100
Item 4	14	100	100	90	76
Item 5	57	22	73	10	41
Item 6	86	100	91	30	77
Item 7	71	22	64	30	47
Item 8	71	33	55	60	55

Tabel 5 Percentage dat het item als kernitem beoordeelt in station 4.

	Stafleden	Observatoren	Deelnemers	Simulatieaios	Gem. 4 groepen
Item 1	86	89	90	40	76
Item 2	71	33	64	30	50
Item 3	100	100	100	100	100
Item 4	100	33	91	80	76
Item 5	71	56	73	20	55
Item 6	57	44	27	10	35

Tabel 6 Aantal deelnemers met aantal kernitems voldoende/goed.

	Station 1	Station 2	Station 3	Station 4
1 kernitem voldoende/goed	1	2	1	0
2 kernitems voldoende/goed	4	2	0	4
3 kernitems voldoende/goed	13	14	3	14
4 kernitems voldoende/goed			4	
5 kernitems voldoende/goed			10	

De tabellen 2 tot en met 5 geven het percentage respondenten weer per groep dat het betreffende item na de tweede ronde als kernitem beoordeelt. In de laatste kolom is het gemiddelde van de vier groepen weergeven.

Wanneer bij het bepalen van het afkappunt gerekend wordt met 75% overeenstemming (gemiddeld over de vier groepen), blijkt dat van de in totaal 25 items er 14 als kernitem worden geïdentificeerd. Worden nu de scores van de deelnemers op de toets van najaar 2004 met deze norm benaderd, dan blijkt dat op de verschillende stations 56-78% van de deelnemers alle kernitems heeft gehaald en dus de vaardigheid volgens deze norm op voldoende niveau beheerst (zie tabel 6). Slechts 4 van de 18 deelnemers hebben op alle stations alle kernitems gehaald.

Conclusie en beschouwing

Het blijkt met deze procedure mogelijk een norm te stellen, die in elk station leidt tot een aanvaardbare verdeling van voldoendeonvoldoende.

Als de – op zich arbitraire – 75% grens als consensusafkappunt wordt genomen, zijn in station 1 drie van de vier items kernitems. Voor station 2 geldt dat voor drie van de zeven items, voor station 3 voor vijf van de acht items, en voor station 4 voor drie van de zes items. Met de uitkomst dat 56-78% van de deelnemers in een recente toets de stations haalt, lijkt deze norm acceptabel en te passen bij het educatieve doel.

Elke opleider die nu aan deze toets meedoet, kan zich aan de norm spiegelen en beseft dat het onvoldoende scoren op één van de kernitems leidt tot een onvoldoende score op het gehele station. Dat is in educatieve zin een belangrijk gegeven: het gewenste opleidersgedrag is zo nog steviger neergezet. Daar kan de deelnemer zich in de voorbereiding op de toets en in zijn scholingstraject na de toets specifiek op richten.

Het stellen van deze norm is een experiment. Geen van de andere auteurs heeft dit tot nu toe gedaan, zo blijkt uit het literatuuroverzicht. Wij zullen de komende tijd ervaring opdoen met het toepassen van de kernitems in de PACT en ze bij de voorbereiding op de toets aan de deelnemers en observatoren bekend maken.

In de nabije toekomst zullen meer stations worden ontwikkeld, waarbij mogelijk van meet af aan al aan een kernitem-consensusprocedure kan worden gewerkt.

Dankwoord?Dank aan dr. J.J.M. Janssen en prof. dr. S. Schol voor hun medewerking aan het tot stand komen van dit artikel. Veel dank aan drs. H.E.P. Bosveld voor zijn methodologische inbreng.

Summary

Introduction: In the postgraduate General Practice Programme of the Department of Family Medicine in Groningen, the Netherlands, a skills test involving multiple stations is used to assess the competence in teaching of General Practice trainers. Trainers who participated in the test were interested to learn whether their performance on (parts of) the test was adequate. We sought to arrive at a standard that was suited to the formative purpose of the test and we examined the consequences of selecting a certain standard.

Method: A consensus procedure was conducted among 7 staff members, 9 observers, 11 trainers who participated in the test and 11 general practice trainees to identify key items for each of four test stations. Consensus among at least 75% of the participants was considered critical. We calculated for each station the percentage of participants who had satisfactory/good scores on all the key items.

Results: The consensus procedure identified 14 of the 25 items in the four stations as key items. Based on the results of a recent test each station would result in a pass for 56-78% of the participants. Conclusion and discussion: These results seem to justify the conclusion that it is possible to set an acceptable standard that is compatible with the formative goals of the test. (Lettinga JP, Boendermaker PM. An acceptable standard for a test of GP trainers' teaching skills: Did I do well enough? Dutch Journal of Medical Education 2007;26(2):75-81.)

Literatuur

SVUH-werkgroep kwaliteit opleiders. Een systeem voor toetsing van huisartsopleiders. Utrecht: SVUH; 2004.

Schol S. Een agogische vaardigheidstoets voor huisartsenpraktijkopleiders: beschrijving van een nieuw instrument. Tijdschrift voor Medisch Onderwijs 2000;19(5):179-86.

Schol S. A multiple-station test of the teaching skills of general practice preceptors in Flanders, Belgium. Acad Med 2001;76 (2):176-80.

Prislin MD, Fitzpatrick C, Giglio M, Lie D, Radecki S. Initial experience with a multi-station objective structured teaching skills evaluation. Acad Med 1998;73(10):1116-8.

Morrison EH, Boker JR, Hollingshead J, Prislin MD, Hitchcock MA, Litzelman DK. Reliability and validity of an objective structured teaching examination for generalist resident teachers. Acad Med 2002;77(10):s29-s32.

Boendermaker PM. Meesterschap. Van verkenning naar herkenning van de goede huisartsopleider [dissertatie]. Maarssen: Elsevier; 2003.

Zabar S, Hanley K, Stevens DL, Kalet A, Schwartz MD, Pearlman E, et al. Innovations in education and clinical practice. Measuring the competence of residents as teachers. J Gen Intern Med 2004; 19(5):530-3.

Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995;311:376-80.